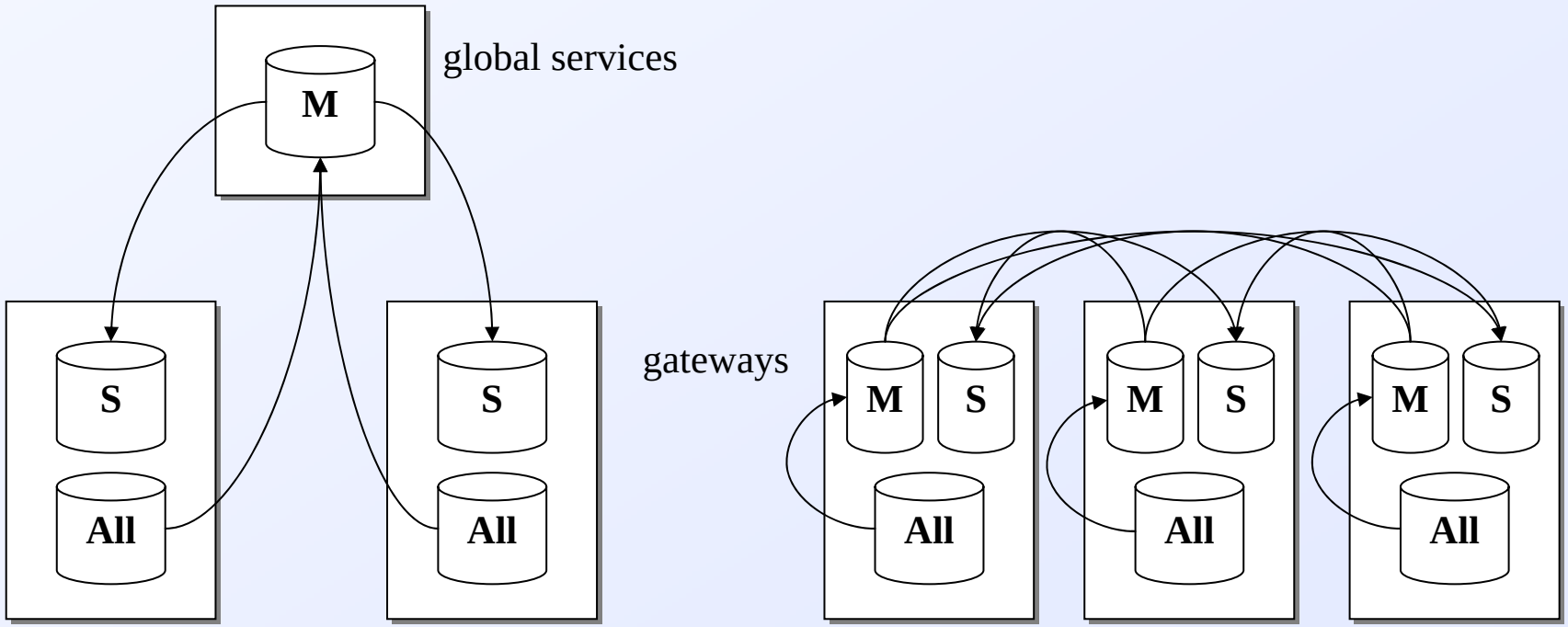
Federated Metadata

- Gateways maintain separate (relational) metadata databases
- How much metadata information should be shared among gateways?
- Search metadata (RDF)
  - Harvested from each gateway and shared with other gateways
- Additional information to share:
  - VO-level attribute information, policies
  - Other?
- We assume it is not scalable to share full metadata among all gateways
  - What is the total volume of metadata?
  - Example: LIGO, hundreds of millions of files, eventual scalability problems

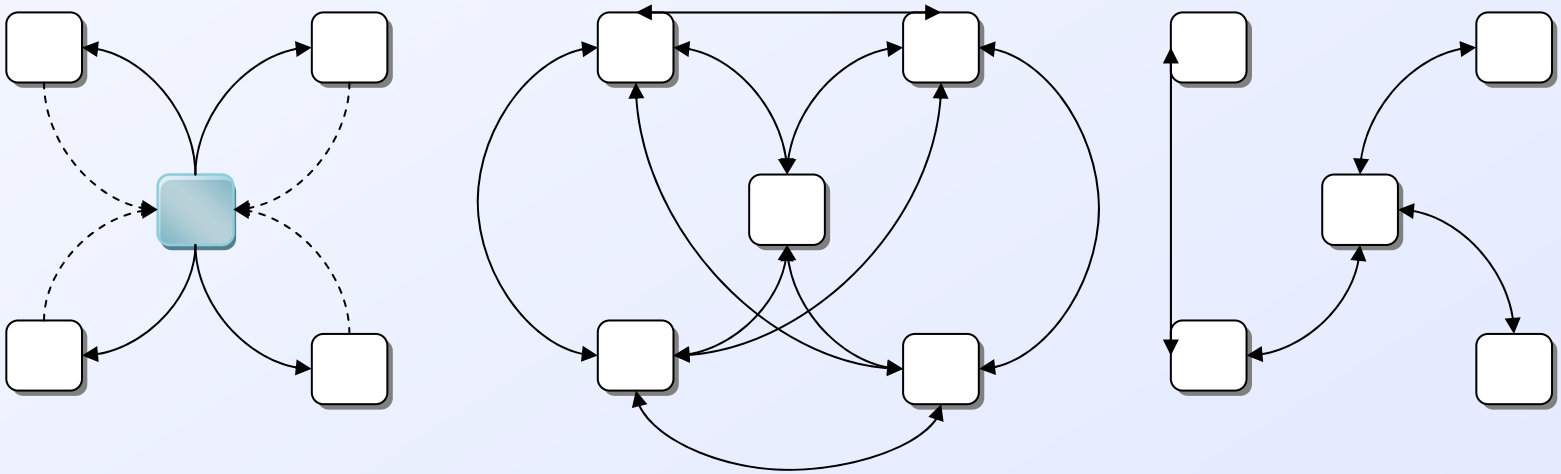
Federated Metadata Issues

- Single master: a global metadata service
- Multiple masters: each gateway acts as master for its local data sets
  - All: the full collection of metadata
  - M: the master copy of search metadata
  - S: the secondary copy of search metadata



Update Topology

- Single master: receives updates from all gateways, sends updates to all gateways
- Multiple masters: fully connected or sparsely connected



Update Issues

- Update frequency
  - More frequent updates provide greater consistency, but consume more bandwidth
- Full vs. incremental updates
  - Incremental updates save bandwidth, can be sent more frequently
- Soft or Hard State
  - Soft state expires and must be refreshed
    - Eliminates or reduces the need to propagate deletes, since entries will expire
    - Requires that full contents must be sent periodically
  - Hard state
    - State is kept until it is explicitly removed

Update Issues

- Size of updates
  - How large will they be?
- Compression
  - Do we need to compress updates?
  - Bloom filter compression (used in RLS) is a lossy compression scheme
    - Perform a series of hash functions, set corresponding bits in a bit map
    - E.g., could hash the RDF triples and set the corresponding bits
    - Searches would be based on the same triple being provided in the query
    - Cannot reconstruct original triples from the Bloom filter
  - May need lossless compression schemes

Metadata harvesting

- OAI Protocol for Metadata Harvesting (OAI-PMH)
  - Will this be used to harvest search metadata?
  - “harvesters” (the client) request updated metadata from “repositories” (the server)
  - Each ESG gateway might setup an OAI repository for their search metadata
  - Other sites (gateways or the global service) would run a harvester to collect metadata updates